

Predicción de niveles de NO₂ en Madrid

Luz Frias

July 17, 2016

Contents

Introducción	2
Normativa	2
Zonas	2
Niveles de actuación	2
Medidas	2
Objetivo del proyecto	3
Datos	3
Contaminación	3
Mediciones históricas	4
Mediciones en tiempo real	4
Información de estaciones	4
Variables predictoras	4
Introducción	4
Metereología	4
La inversión térmica	5
Niveles de tráfico	8
Calendario laboral	8
Análisis descriptivo	8
Distribución de los valores	8
Predicción	14
Valores de NO ₂ con Gradient Boosting Trees	14
Valores de NO ₂ con Elastic Net	16
Valores de NO ₂ con series temporales bayesianas	18
Clasificación de nivel de aviso con Random Forest	22
Próximos pasos	23

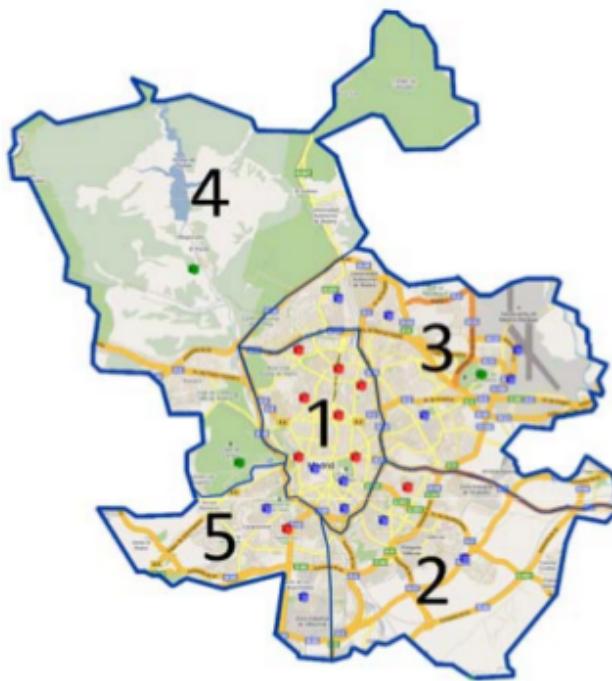
Introducción

Durante el último año, el tema de la contaminación en Madrid ha adquirido gran protagonismo. En parte por la creciente preocupación por el medio ambiente y el efecto de la contaminación en la salud pública, y en parte por las recientes restricciones de circulación impuestas por el Ayuntamiento de Madrid para disminuir los niveles más altos.

Normativa

Zonas

Se establecen 5 zonas en Madrid con estaciones de medición:



Niveles de actuación

Se establecen tres niveles de actuación en función de las concentraciones de NO₂ que se registren en las zonas que se han definido

- Preaviso: cuando en dos estaciones cualesquiera de una misma zona se superan los 180 microgramos/m³ durante dos horas consecutivas.
- Aviso: cuando en dos estaciones cualesquiera de una misma zona se superan los 200 microgramos/m³ durante dos horas consecutivas.
- Alerta: cuando en tres estaciones cualesquiera de una misma zona (o dos si se trata de zona 4) se superan los 400 microgramos/m³ durante tres horas consecutivas

Medidas

Se aplican las siguientes restricciones en cada uno de los escenarios definidos:

- Escenario 1: 1 día con superación del nivel de preaviso
 - Reducción de la velocidad a 70 km/h en la M-30 y accesos
- Escenario 2: 2 días consecutivos con superación del nivel de preaviso ó 1 día con superación del nivel de aviso
 - Reducción de la velocidad a 70 km/h en la M-30 y accesos
 - Prohibición del estacionamiento de vehículos en las plazas y horario del Servicio de Estacionamiento Regulado (SER) en el interior de la M-30
- Escenario 3: 2 días consecutivos con superación del nivel de aviso
 - Reducción de la velocidad a 70 km/h en la M-30 y accesos
 - Prohibición del estacionamiento de vehículos en las plazas y horario del SER en el interior de la M-30
 - Restricción de la circulación en el interior de la almendra central (área interior de la M-30) del 50% de todos los vehículos
- Escenario 4: 3 días consecutivos de nivel de aviso o 1 día de nivel de alerta
 - Reducción de la velocidad a 70 km/h en la M-30 y accesos
 - Prohibición del estacionamiento de vehículos en las plazas y horario del SER en el interior de la M-30
 - Restricción de la circulación en el interior de la almendra central (área interior de la M-30) del 50% de todos los vehículos
 - Restricción de la circulación por la M-30 del 50% de todos los vehículos
 - Restricción de la circulación de taxis libres, excepto Ecotaxis y Eurotaxis, en el interior de la almendra central (área interior de la M-30)

Objetivo del proyecto

Hasta ahora, estas restricciones se aplican de manera reactiva, es decir, tras haber alcanzado niveles altos de contaminación. Este proyecto está motivado por intentar abordar el problema de forma proactiva, prediciendo con al menos un día de antelación niveles de alerta de NO₂.

Datos

Contaminación

Los datos de contaminación los publica el ayuntamiento en su portal de datos abiertos. Los datos se pueden encontrar con dos periodicidades diferentes (horaria y diaria) y en formato histórico o tiempo real (datos del día actual, con actualización horaria y un retraso de aproximadamente una hora).

En este proyecto me he centrado en el estudio de los datos a nivel horario. Algo curioso es que no hay problemas en encontrar el histórico desde 2001 hasta el mes previo al actual, ni los datos de hoy, pero no se publican en ningún lado los datos del mes actual hasta el día previo a la consulta.

El parseo de los ficheros resulta un poco incómodo, ya que, aunque contienen la misma información el histórico y los datos en tiempo real, tienen formatos diferentes:

- El histórico se guarda en un fichero de anchos fijos
- El de datos en tiempo real, en un fichero separado por comas

Además que en cada fila se representa un día, y los datos horarios están en formato ancho (un valor horario en cada columna). Por conveniencia para el entrenamiento, el proceso de limpieza lo pasa a formato largo (una fila por hora).

Mediciones históricas

Se encuentran aquí zipeados por año. Para facilitar reprocesos (durante el análisis detecté que faltaban algunos meses, y la mayoría de ellos han sido repuestos tras avisar al administrador del portal) y la inclusión de nuevos datos, los datos se recogen mediante un proceso que:

- Descarga los zip
- Los descomprime
- Los lee y pasa los procesos de limpieza necesarios
- Los guarda como un fichero por año

Mediciones en tiempo real

Se encuentran aquí. Se ha desarrollado un proceso que:

- Lo lee y pasa los procesos de limpieza necesarios
- Lo guarda como un fichero por día

Información de estaciones

Se encuentra aquí en formato excel. Lo he transformado en un proceso semi-manual a texto plano, transformando las coordenadas de formato grados, minutos y segundos a formato decimal.

Variables predictoras

Introducción

El primer paso fue estudiar cuáles son las variables que influyen en la varación de los niveles del NO₂. Tras algo de documentación en páginas relacionadas y papers científicos, se determina que son:

- Las fuentes principales son las emisiones de vehículos e industriales
- La variación del nivel está relacionada con factores metereológicos, especialmente:
 - Velocidad del viento
 - Humedad relativa
 - Temperatura

Metereología

Curiosamente, una de las cosas que más me ha costado encontrar. Agencias como la AEMET proveen datos históricos a pocos días atrás. Otras páginas te permiten consultar solo algunos datos históricos (temperatura, lluvia y poco más) a través de visualizaciones en su web.

La falta de datos en este ámbito ha causado incluso a que algunos particulares vendan la información, como p.ej. datosclima.es.

Finalmente los datos utilizados para este proyecto se han escapado de OGIMET, que proporciona registros diarios y horarios para un conjunto reducido de estaciones. He escogido la estación de Madrid - Barajas por ser la más cercana a la ciudad de entre las disponibles.

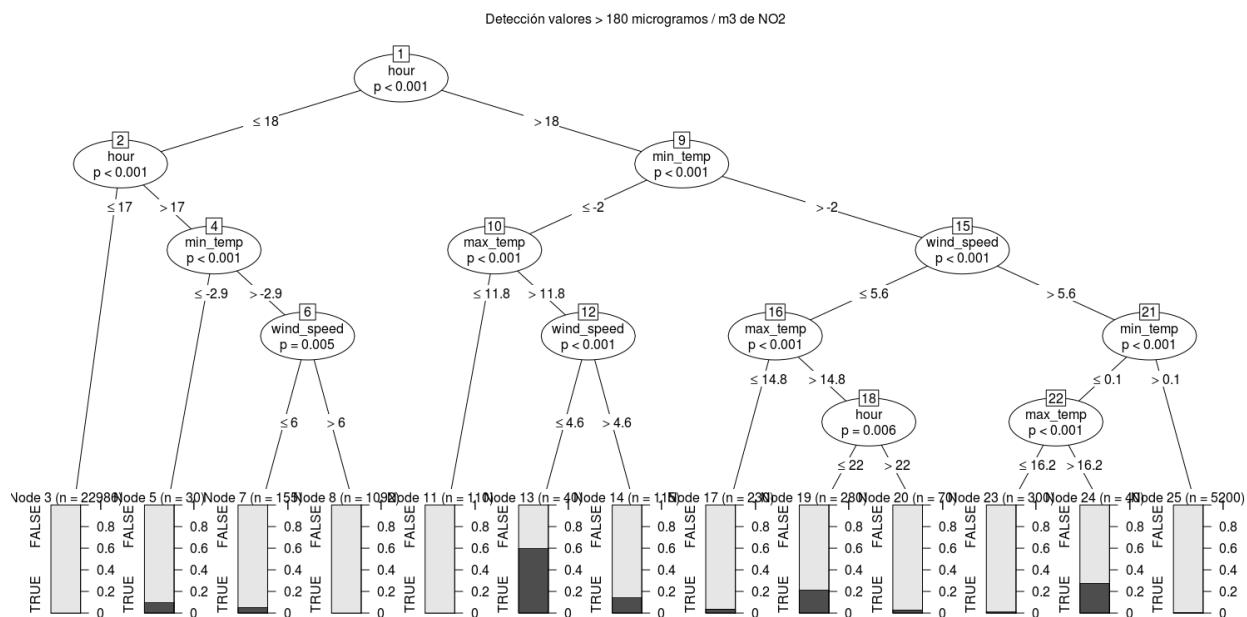
En el scraping se extraen:

- A nivel diario:
 - Temperatura media, mínima y máxima
 - Humedad relativa
 - Velocidad del viento
 - Nivel de precipitaciones
- A nivel horario:
 - Temperatura

La inversión térmica

Un descubrimiento importante durante la construcción de los modelos predictivos fue el efecto de la inversión térmica (que hasta ese momento no conocía).

En el siguiente árbol se puede observar cómo, en días con temperaturas mínimas bajas y máximas altas, aumenta considerablemente la probabilidad de encontrar valores superiores a 180 microgramos / m³ de NO₂



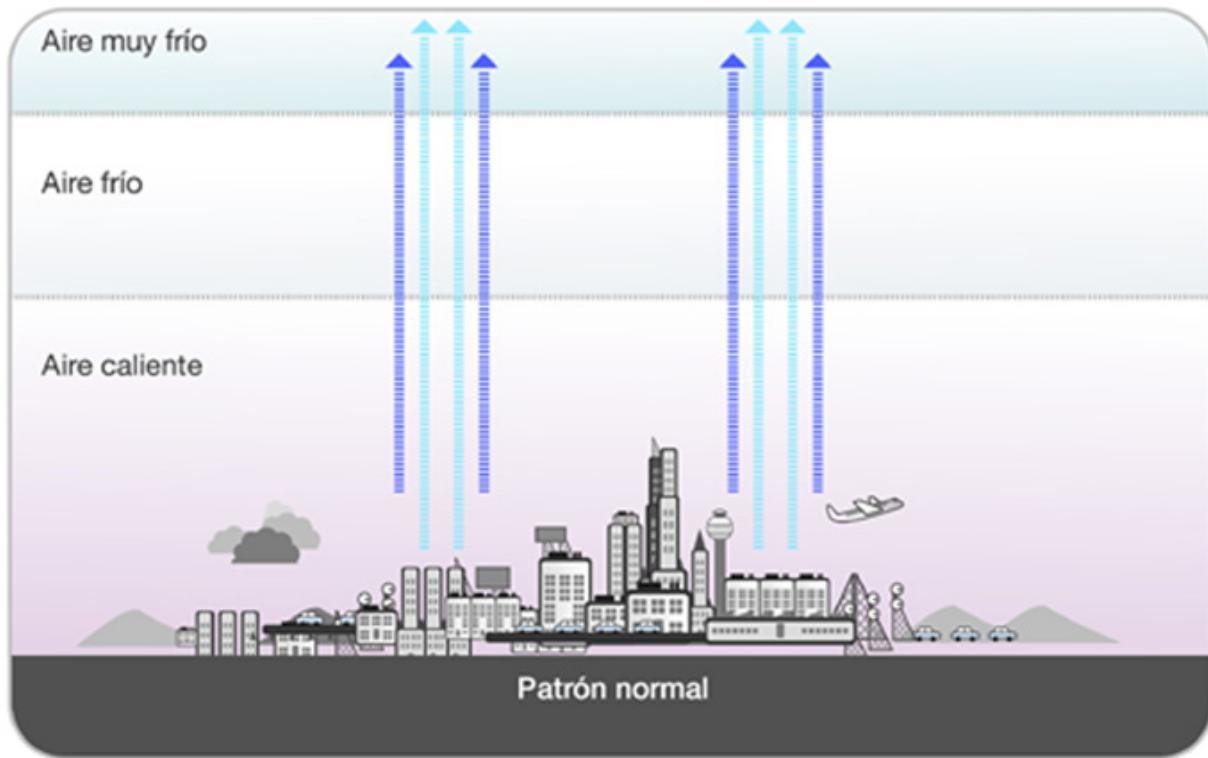
A partir de ese descubrimiento, incluí la diferencia entre la temperatura mínima y máxima entre las variables predictoras.

Pero, ¿qué es la inversión térmica? En situaciones normales, el aire se mueve constantemente y las capas que lo forman suelen ordenarse por su temperatura, con las más frías circulando en la parte alta de la atmósfera y las más calientes, abajo.

Cuando ese ciclo de movimiento se interrumpe, se forma una capa de aire frío que queda inmóvil sobre el suelo e impide la circulación atmosférica. Este fenómeno es la inversión térmica y se produce con más frecuencia en

las noches despejadas de invierno, cuando el suelo ha perdido calor por radiación y las capas de aire cercanas a él se enfrián más rápido que las capas superiores.

Cuando el aire se mueve con normalidad hace circular grandes cantidades de polvo, humo y partículas suspendidas, eliminando la contaminación y limpiando la atmósfera de manera natural. Por eso, cuando la inversión térmica inmoviliza las capas inferiores cercanas al suelo sobre una ciudad, quedan atrapados los contaminantes suspendidos y la población se expone a respirar un aire más contaminado de lo normal.



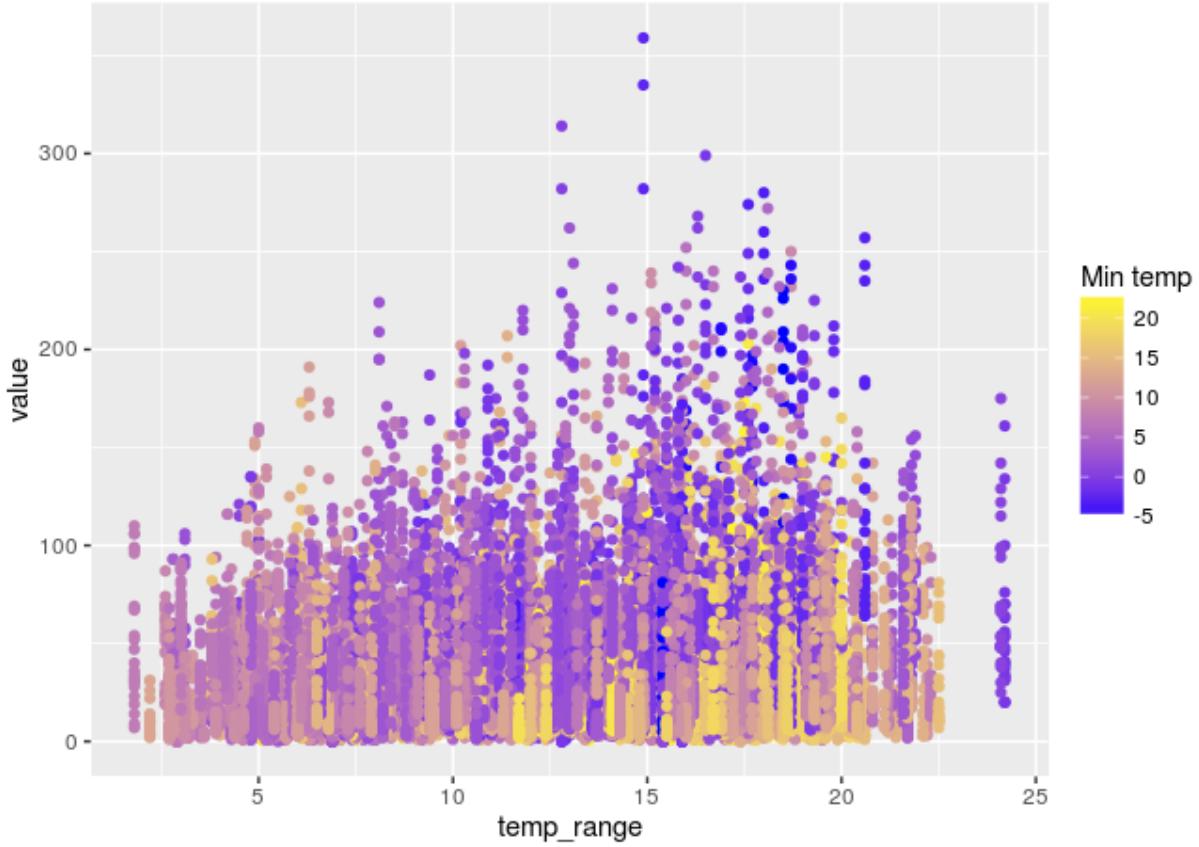
Aire muy frío

Aire caliente de inversión

Aire frío

Inversión térmica

En nuestros datos, si visualizamos los niveles de NO₂ en base a la diferencia de temperaturas mínima y máxima, y la temperatura mínima, podemos observar como un porcentaje importante de los valores altos de NO₂ pueden estar relacionados con bajas mínimas y una diferencia grande con la máxima.



Niveles de tráfico

Aquí se puede encontrar un histórico de los niveles de tráfico en la M-30 desde 2013, que se actualiza diariamente.

Vienen datos totales en formato xml con un nodo por día.

La inclusión de los datos de niveles de tráfico, supone en general una mejora de la predicción, pero “perdemos” el histórico de 2001 a 2013 que sí tenemos disponible en datos de contaminación y metereología.

Calendario laboral

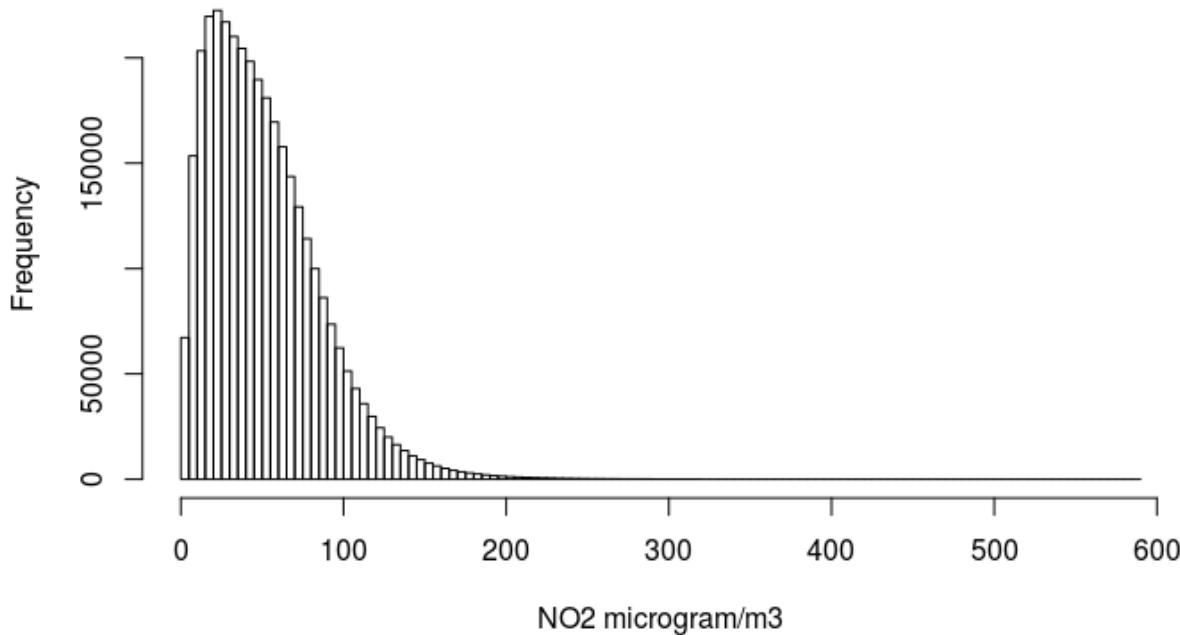
Con el objetivo de poner en producción el modelo, hay que predecir la intensidad de tráfico a futuro. Para ello, he incluido los datos de festivos en la ciudad de Madrid. Ha sido un proceso semi-manual y con algunos datos dudosos en los años más antiguos.

Análisis descriptivo

Distribución de los valores

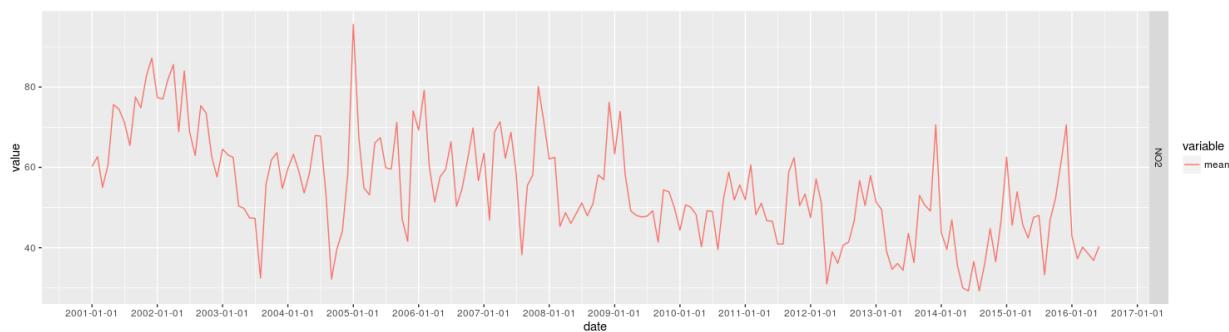
Los valores se concentran en niveles bajos, formando una distribución con sesgo positivo.

Histogram of NO2 values

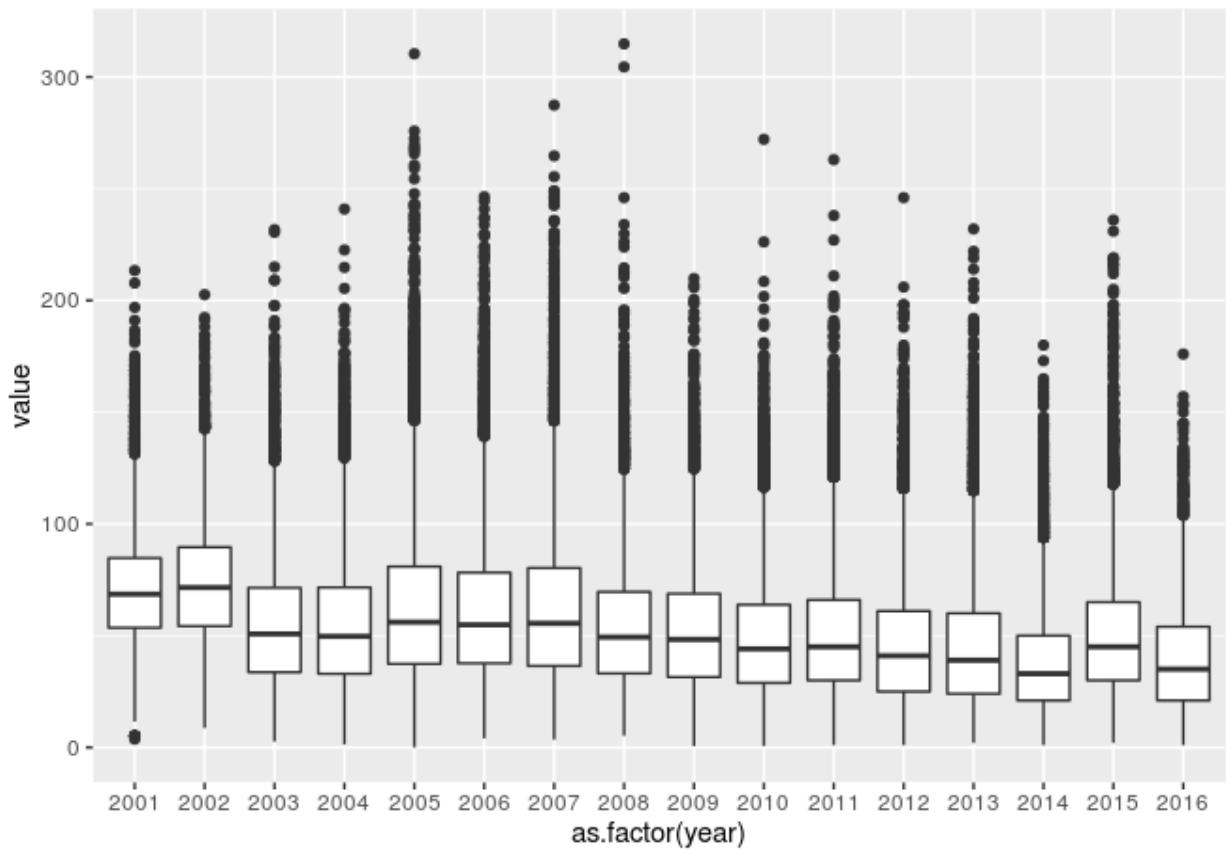


Los siguientes gráficos se han pintado en base a las mediciones de la estación de plaza de España, para observar la evolución y no ensuciar la visualización con estaciones que se han dado de alta y de baja en los diferentes años.

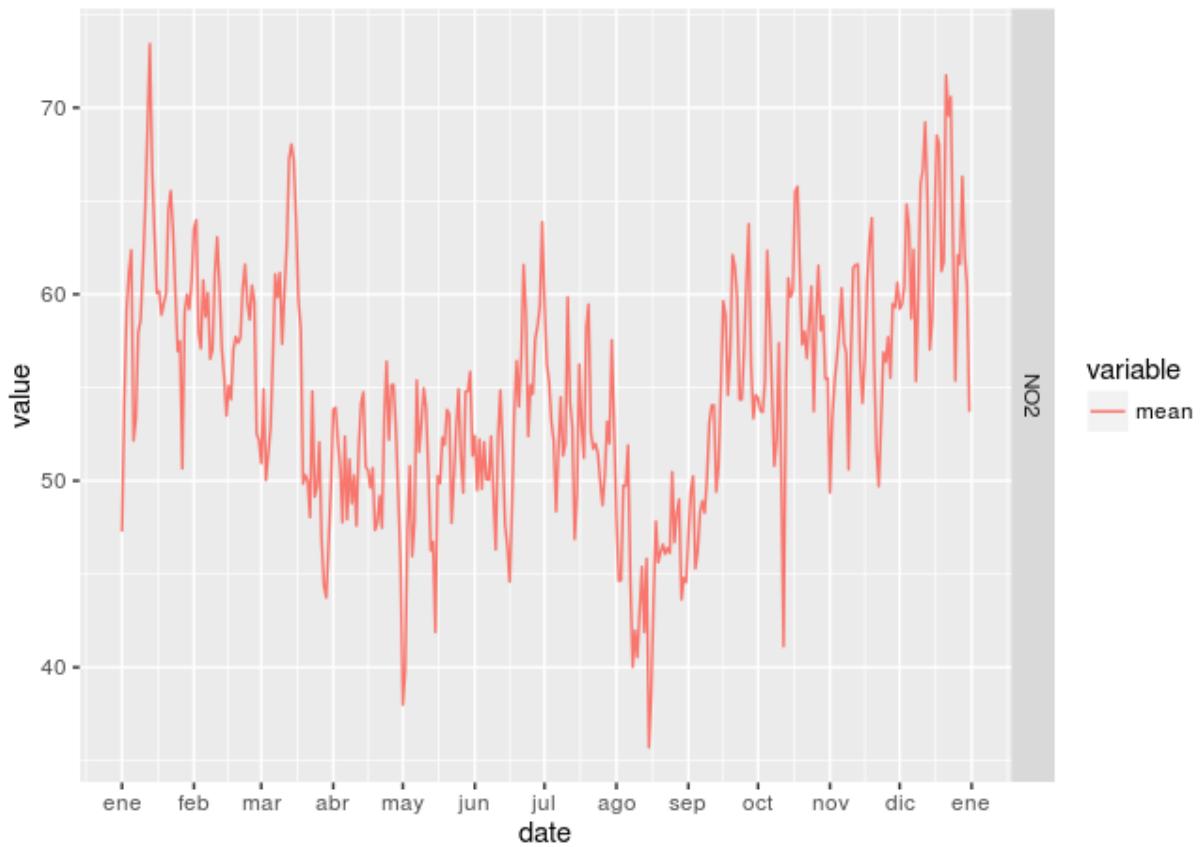
La evolución histórica ha sido de una ligera disminución de la media mensual de valores de NO₂. Pero en los últimos años, los inviernos están registrando medias más altas con respecto al periodo 2010 a 2013.



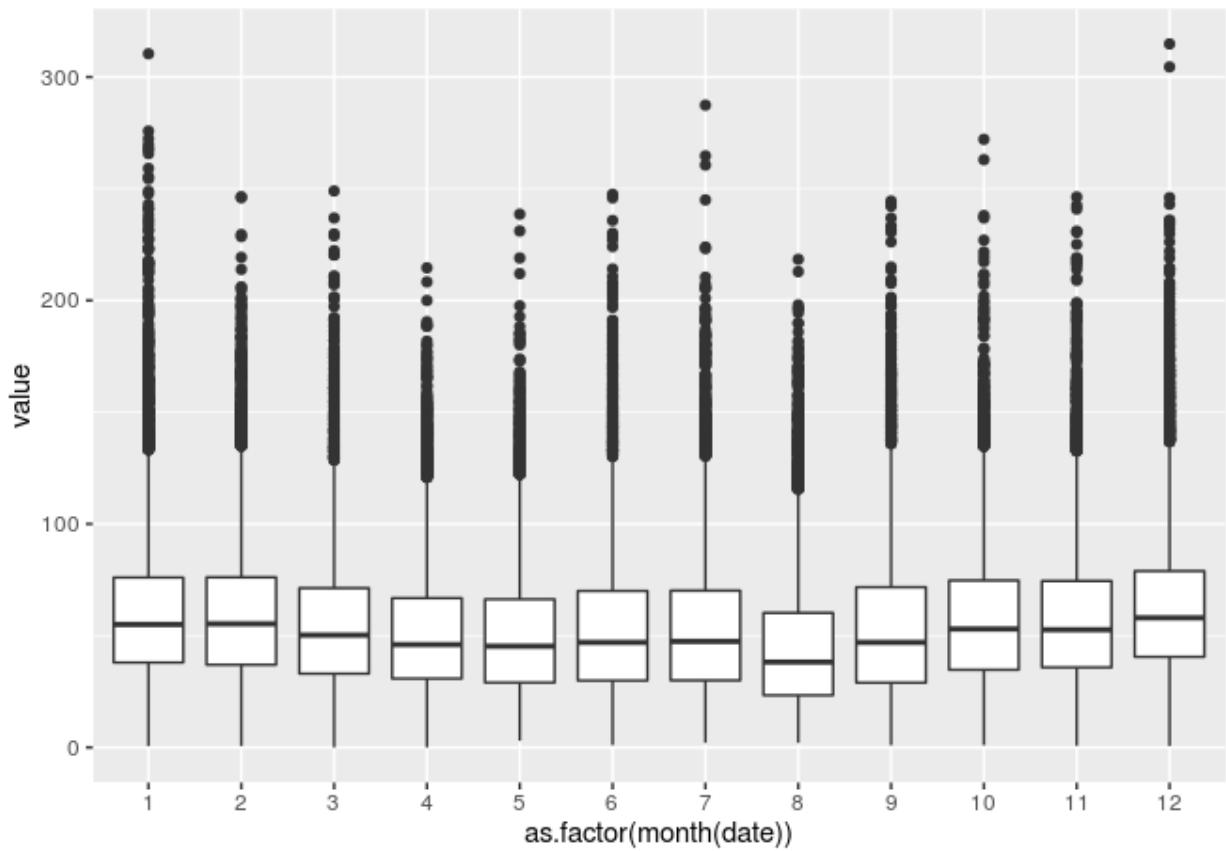
Además se registran valores altos, a menudo superando el nivel considerado de alerta de 200 mg/m³.



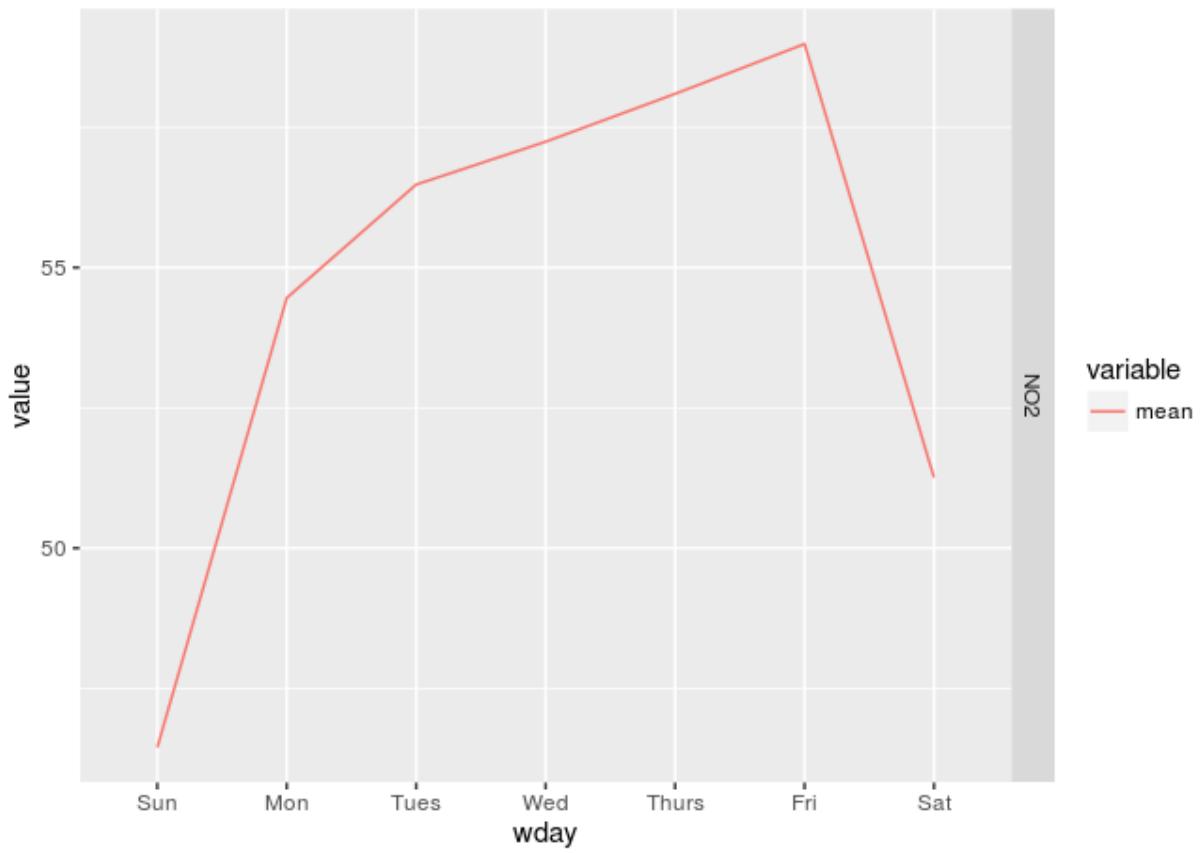
Dentro de cada año, los meses fríos son los que registran las medias más altas.



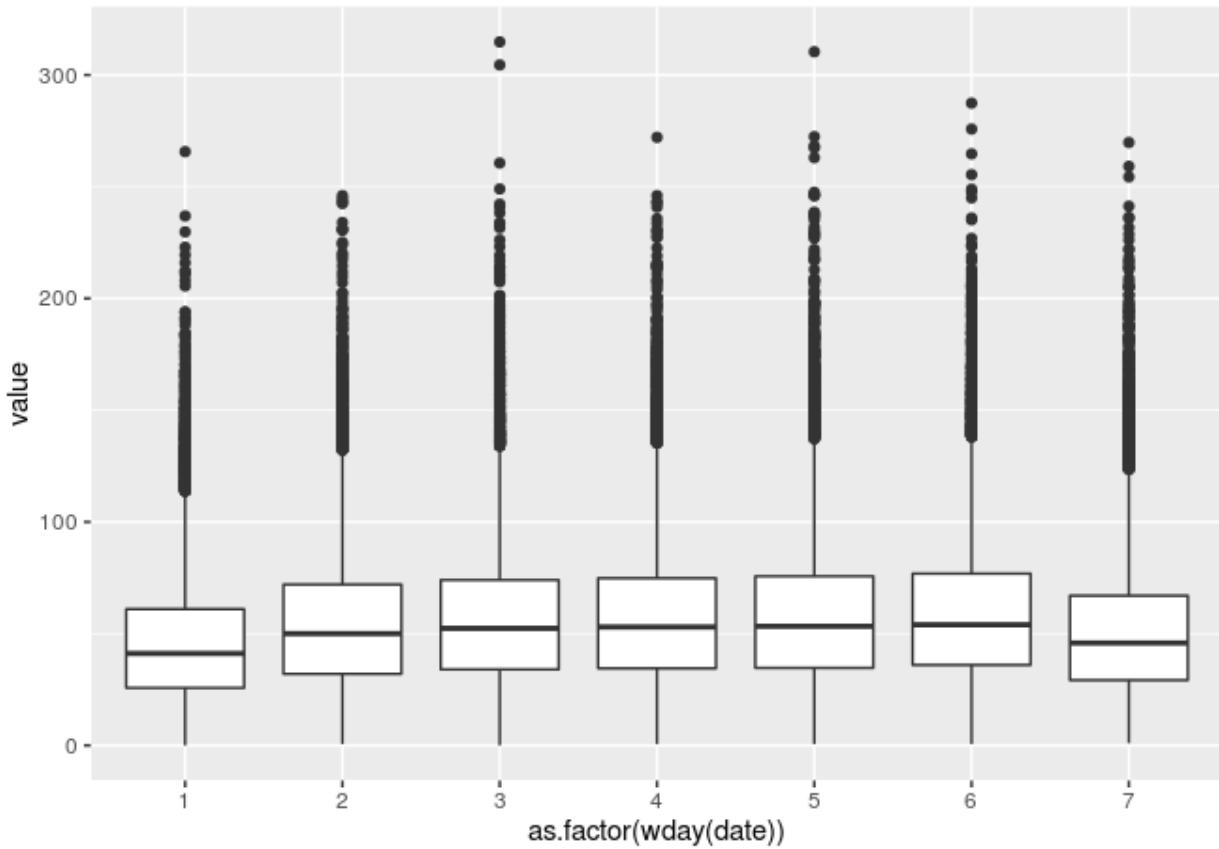
Aunque durante todo el año se registran valores altos de contaminación, son algo más frecuentes y elevados en invierno.



Dentro de una semana, se observan valores más altos en los días laborales con respecto al fin de semana. Además, se observa un posible efecto acumulativo.



Los valores más altos de contaminación se dan en cualquier día de la semana.



Predicción

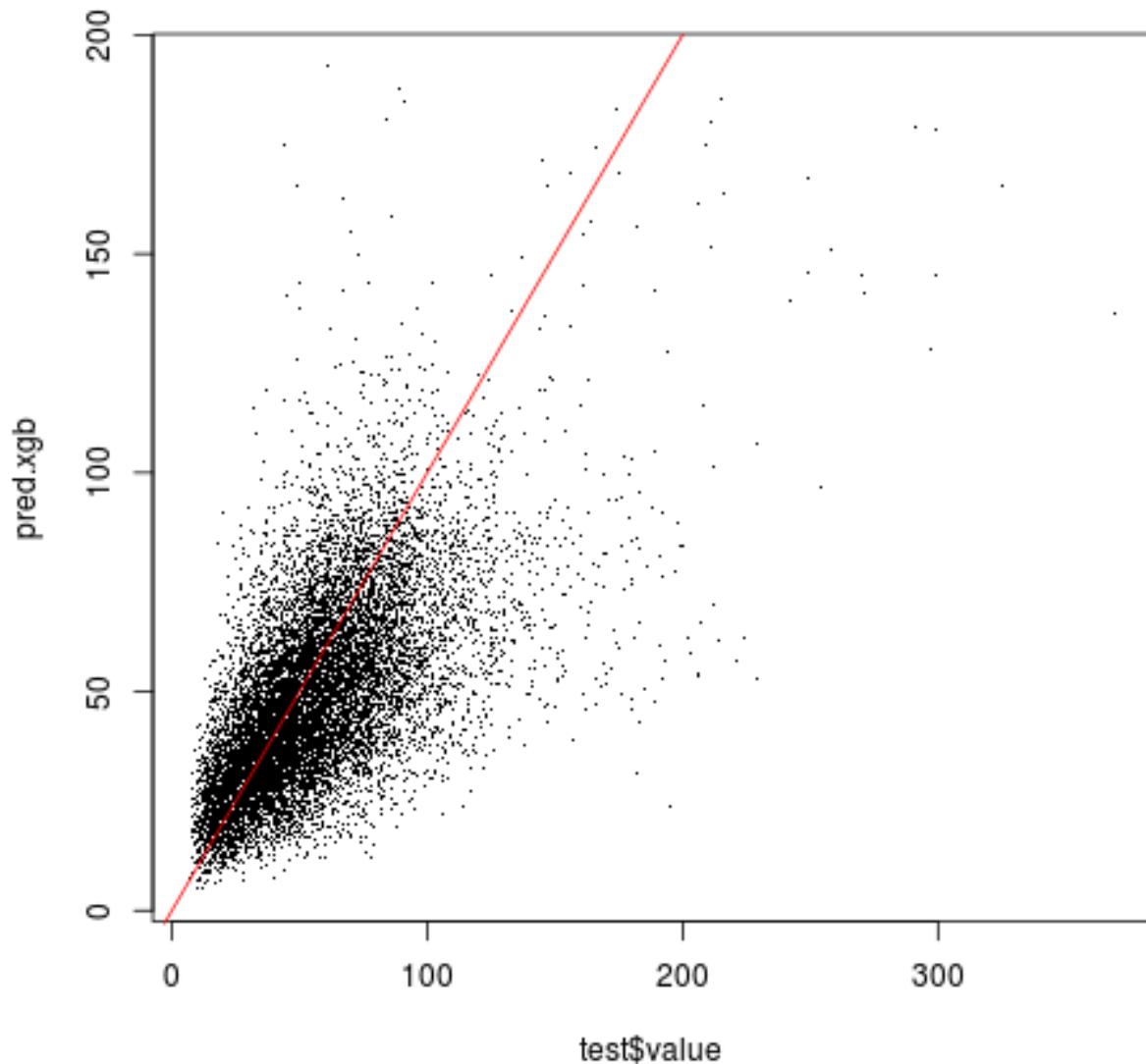
Debido a las características de cada punto de medición, en todos los casos se ha creado un modelo por cada estación. Un ejemplo es cómo influye el nivel de tráfico a estaciones cercanas a puntos de alta densidad de circulación con respecto a puntos situados en grandes zonas ajardinadas (p.ej. parque Rey Juan Carlos I).

Valores de NO₂ con Gradient Boosting Trees

Los modelos realizados tienen las siguientes características:

- Paquete de R xgboost.
- Entrenados mediante validación cruzada.
- Los datos incluidos en cada iteración del entrenamiento durante la validación cruzada es un subconjunto pequeño de los datos (7.5%). Esto es para evitar overfitting, ya que los datos de un mismo día pero diferentes horas tienen muchos datos en común (intensidad tráfico, datos meteorológicos a nivel diario, ...) y con árboles profundos tiende a memorizar los valores fijando los datos comunes.
- Se sobrescribe la métrica de evaluación a minimizar (por defecto RMSE) por una customizada, que penaliza los errores en valores altos. Es decir, es peor equivocarse por 30 microgramos en un valor real de 180 que en uno de 20. Aunque haciendo la prueba de volver a entrenar optimizando el RMSE los resultados son prácticamente los mismos.
- Tuning de parámetros mediante caret.
- Separación de conjunto de entrenamiento y validación por fecha, intentando simular que predecimos valores futuros en base a observaciones pasadas.

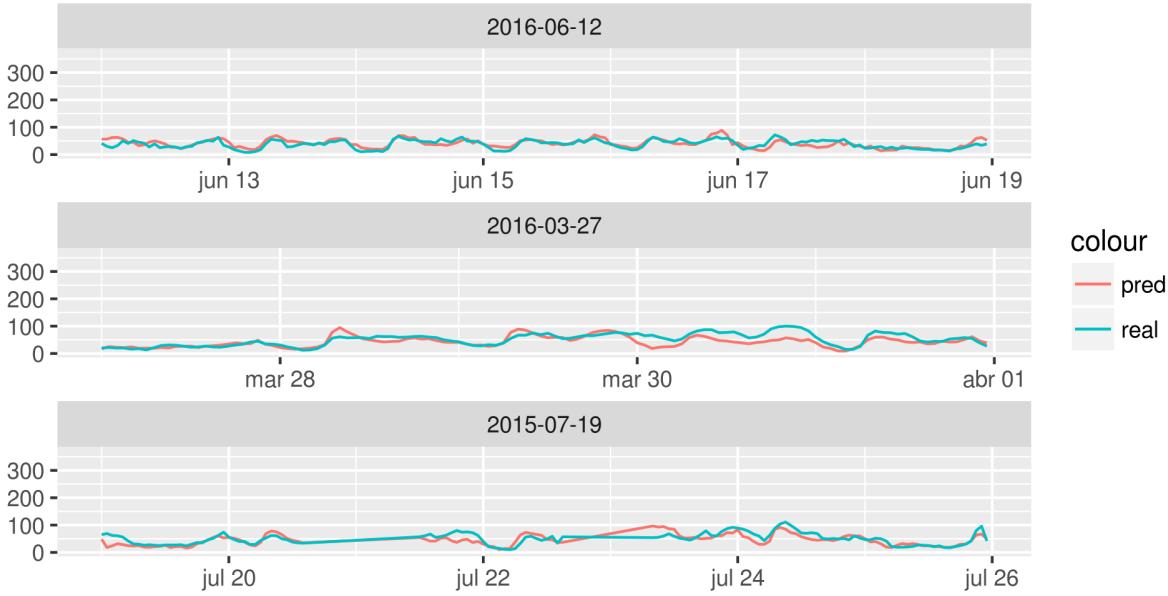
Los resultados son en general buenos para valores normales, pero se pierden a veces los picos. P.ej. esta es la comparación entre valores reales y predichos en la estación de C/Alcalá con C/O'Donell.



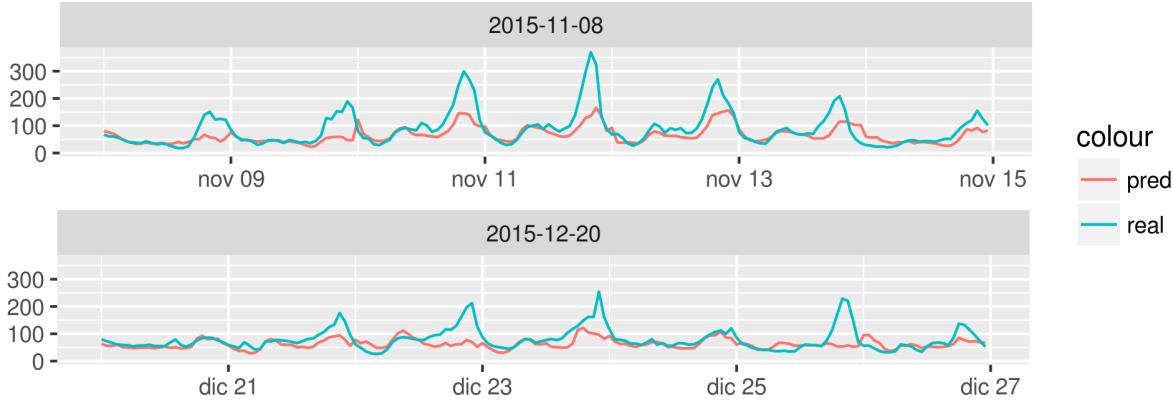
El rendimiento general sobre los datos de validación es:

- MAE: entre 9 - 19
- RMSE: entre 12 - 27

Aquí se puede ver cómo los valores bajos y normales se predicen muy bien:



Y aquí otras no tan precisas:



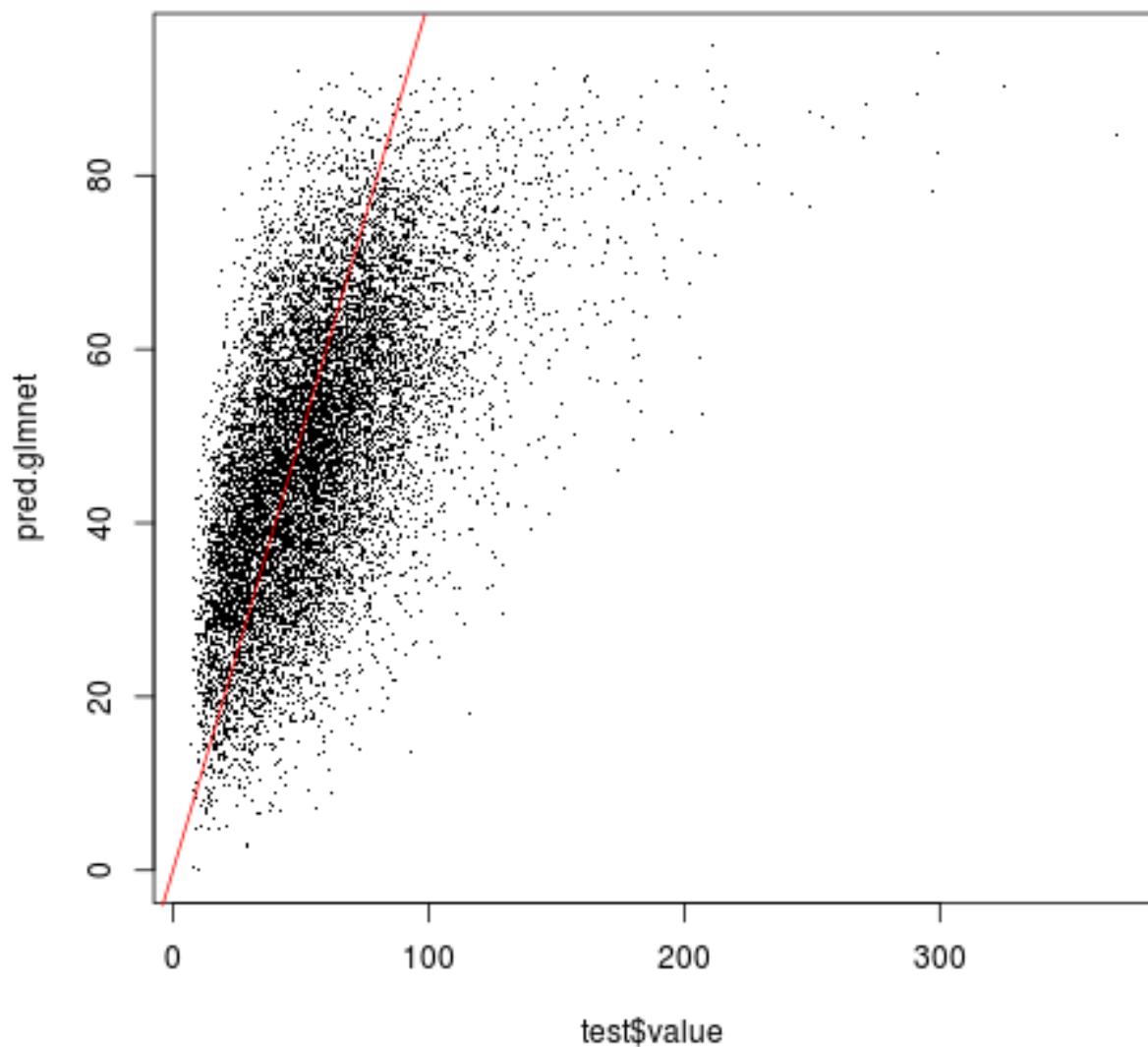
Valores de NO₂ con Elastic Net

Para intentar mejorar la predicción en los picos, vamos a intentarlo cambiando de modelo a uno que, en lugar de utilizar internamente árboles, utilice elastic net (lasso + ridge)

De forma similar al entrenamiento anterior, los modelos realizados tienen las siguientes características:

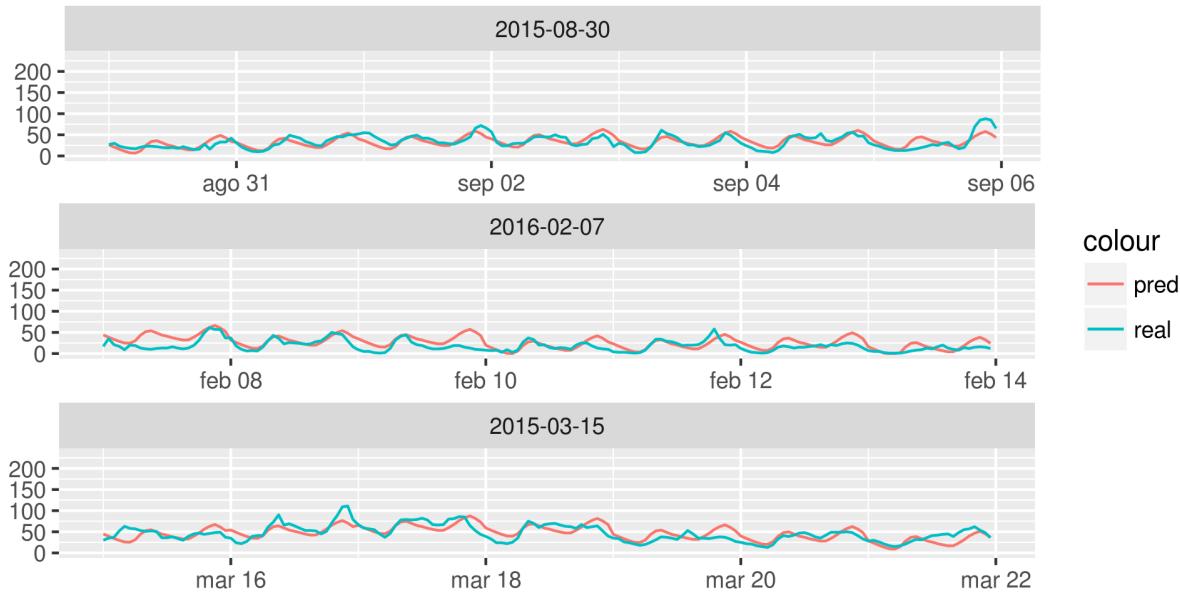
- Paquete de R glmnet.
- Entrenados mediante validación cruzada.
- Tuning de parámetros mediante caret.
- Separación de conjunto de entrenamiento y validación por fecha, intentando simular que predecimos valores futuros en base a observaciones pasadas.

Desgraciadamente, los resultados son lo contrario de lo que esperábamos, capturando peor los picos. P.ej. esta es la comparación entre valores reales y predichos en la estación de C/Alcalá con C/O'Donell (misma validación que en el caso de xgboost).

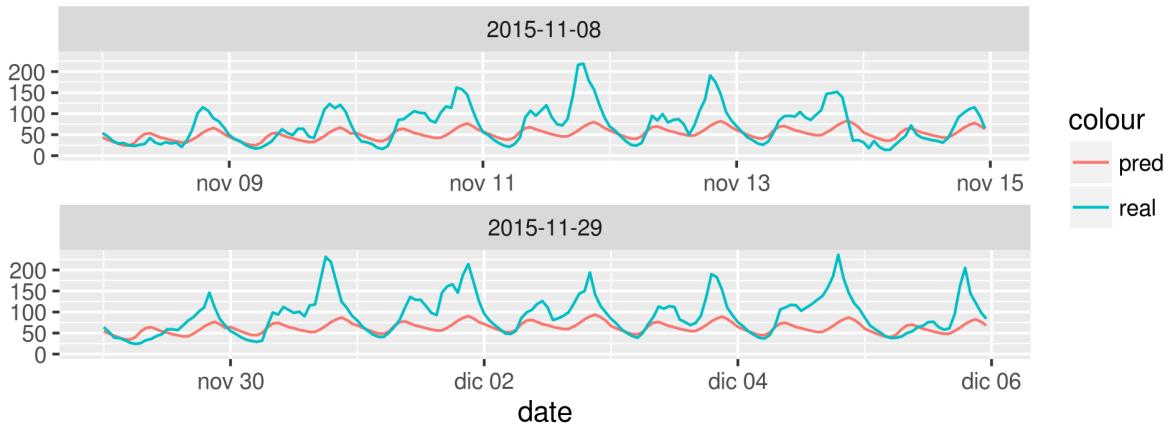


El rendimiento general es algo peor que en el caso anterior.

De forma equivalente, los valores bajos y normales se predicen bien:



Y los picos se pierden:



Valores de NO₂ con series temporales bayesianas

En base a los resultados anteriores, parece que no podemos predecir con exactitud cuándo se va a producir un pico. Pero podemos modificar el problema incluyendo intervalos de confianza, y alertar de aquellos que, con una probabilidad, superen ciertos niveles. Además, nuestros datos son una serie temporal, y hasta el momento no estamos explotando esa estructura interna durante el modelado.

Por ello este entrenamiento lo realizamos utilizando series temporales bayesianas, concretamente el paquete de R bststs. Es lo que utiliza internamente el paquete de Google CasualImpact.

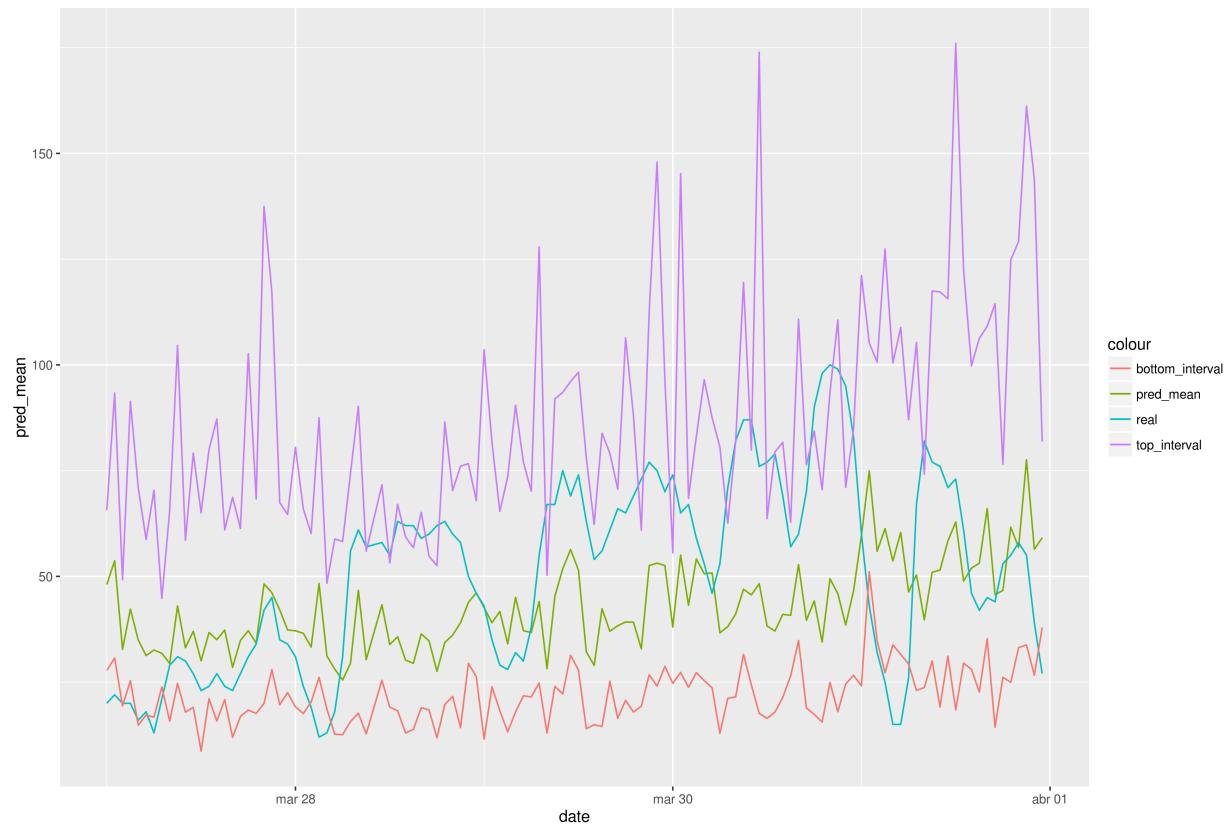
Este modelo proporciona en la predicción distintos valores relacionados con la distribución de la y: la media, mediana y los cuantiles solicitados.

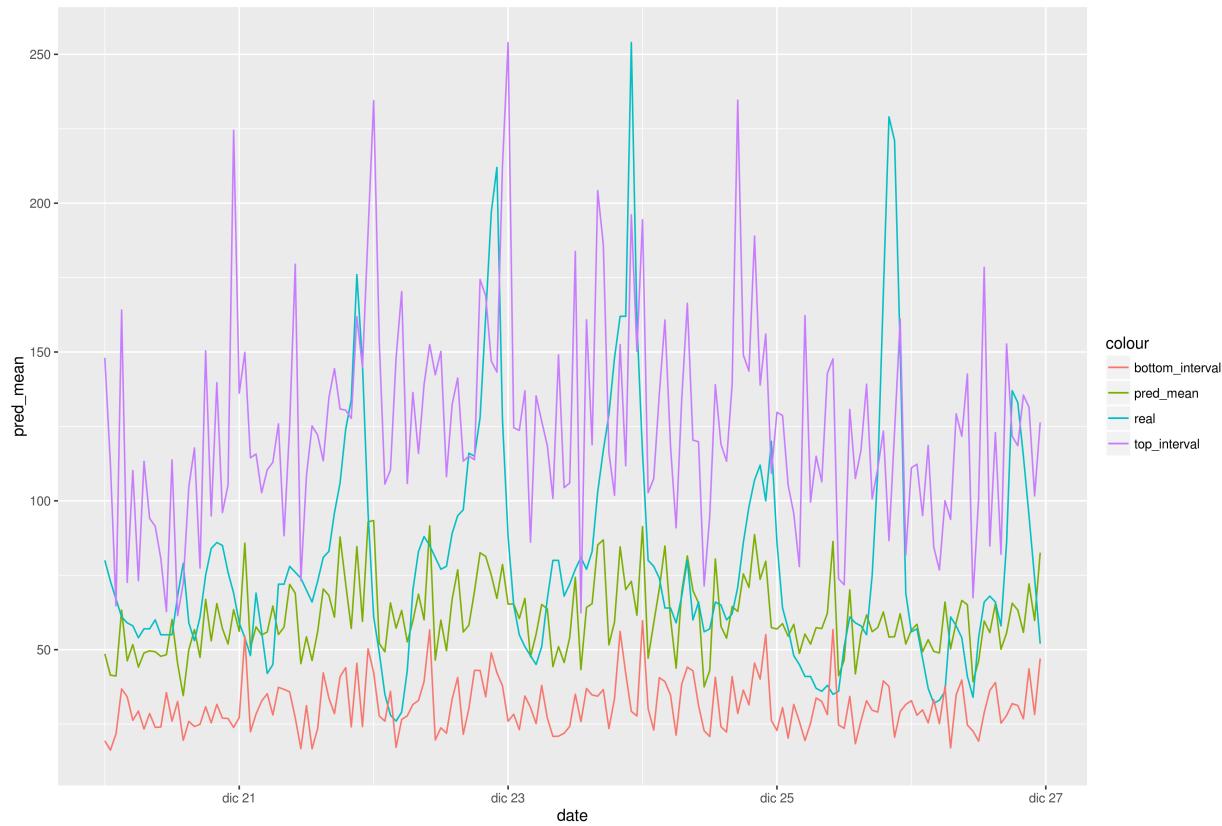
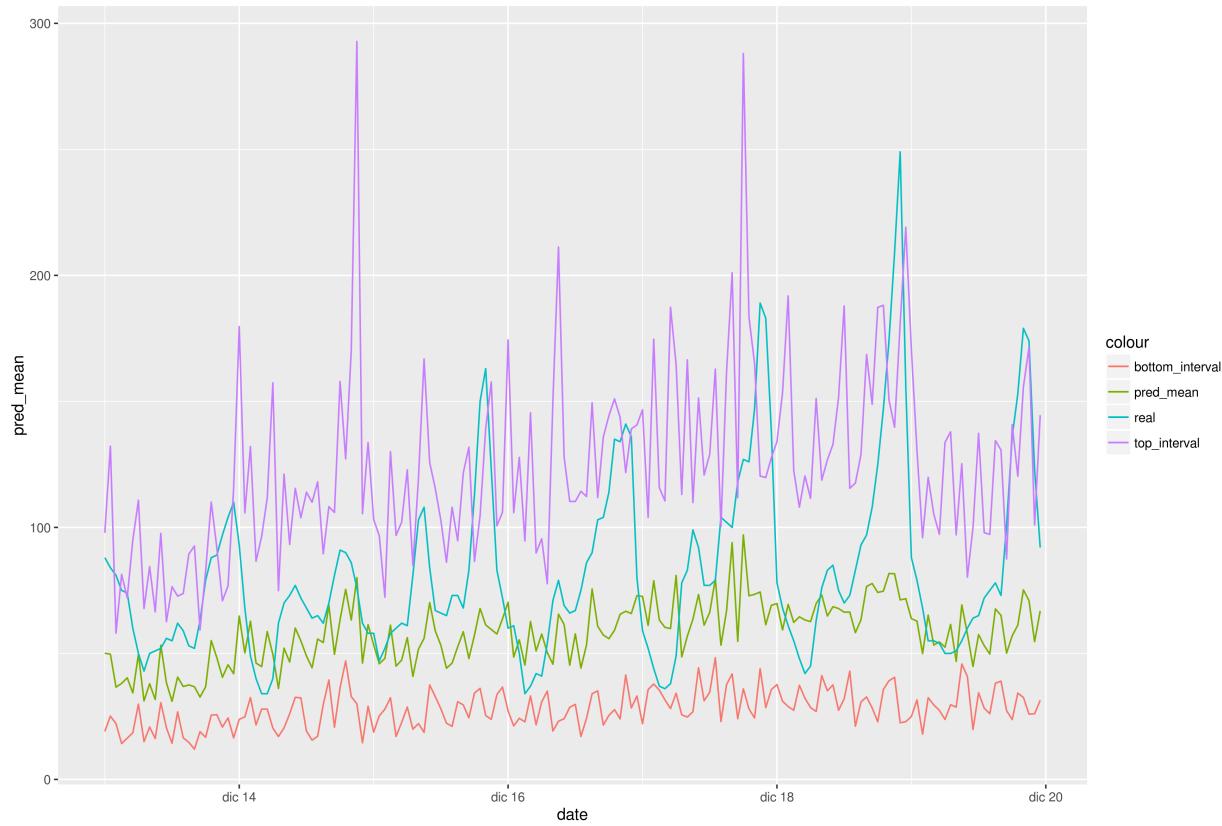
En el entrenamiento le hemos indicado que existe una estacionalidad diaria.

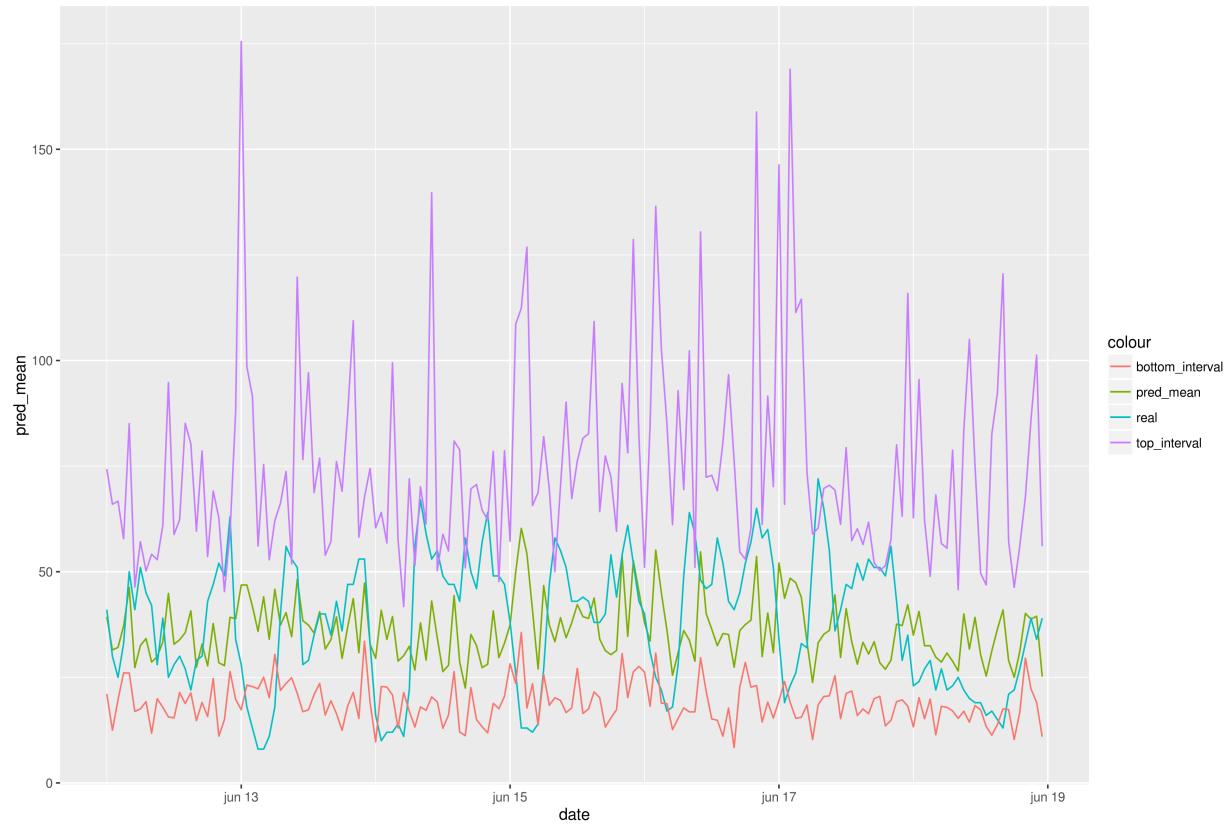
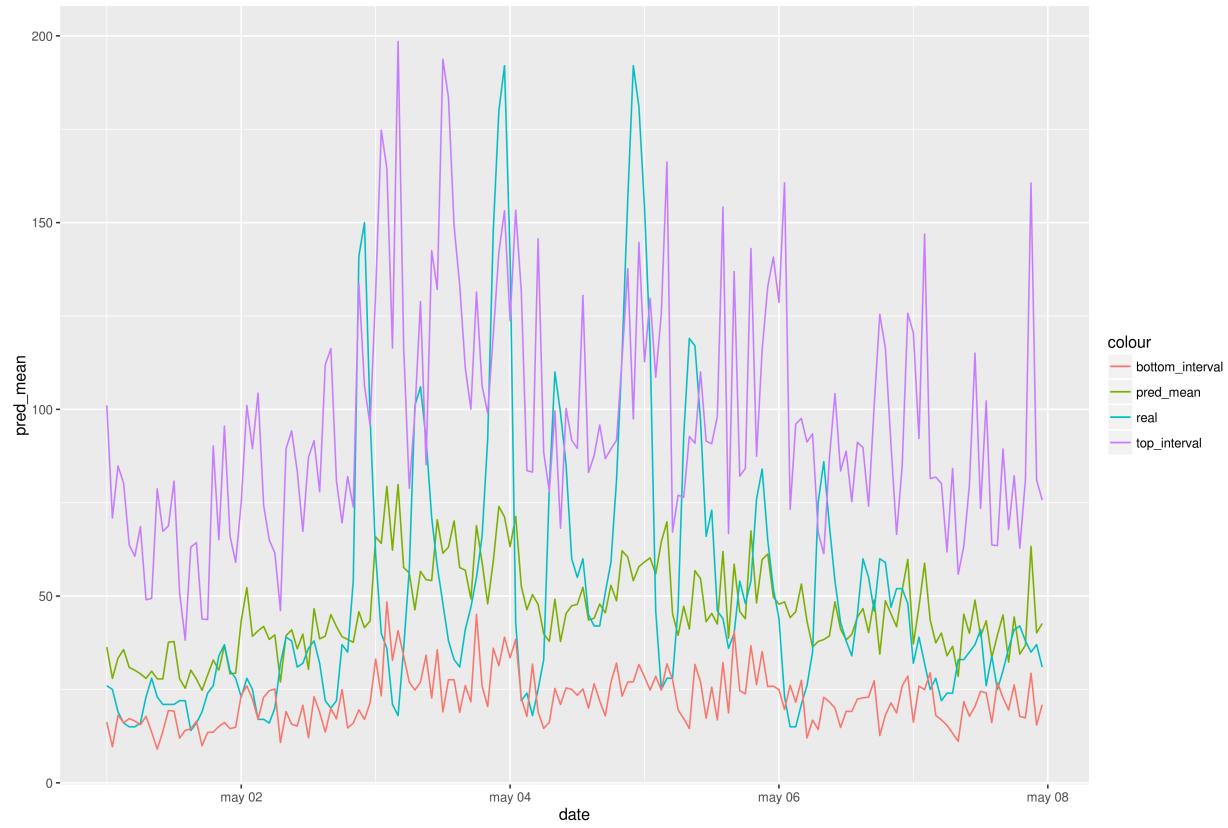
La validación se ha realizado entrenando el algoritmo con datos hasta justo el instante anterior de la fecha que queremos predecir, y prediciendo la semana siguiente.

La precisión del modelo, si comparamos la media de la predicción contra el valor real, es significativamente peor que en los dos casos anteriores. Pero ganamos la información de cómo se puede distribuir esa variable de salida.

Algunos ejemplos de predicción, para la estación de C/Alcalá con C/O'Donell y extrayendo los cuantiles 0.05 y 0.95 son:







En general, los valores reales están por debajo del intervalo superior, y acompañan al valor real (reales más

bajos tienen intervalos superiores también más bajos). Pero se observan casos extremos, como incrementos del intervalo superior con valores bajos del real, o el real superando el intervalo superior.

Clasificación de nivel de aviso con Random Forest

Lo siguiente es una simplificación del problema. En lugar de intentar predecir el valor horario del NO₂, transformamos la pregunta a si en un determinado día se va a superar o no el nivel de preaviso (180 microgramos/m³ durante 2 horas) y de aviso (200 microgramos/m³ durante 2 horas).

Los modelos realizados tienen las siguientes características:

- Paquete de R randomForest.
- Separación de conjunto de entrenamiento y validación por fecha, intentando simular que predecimos valores futuros en base a observaciones pasadas.
- Al contrario que en los modelos anteriores, no usamos el nivel de tráfico, y así podemos entrenar con todo el conjunto de datos.
- Se agrupan los datos a nivel diario.
- Se incluye como variable predictora el percentil 95 de los niveles del día anterior (para tener en cuenta el efecto acumulativo).
- Como el modelo está muy desbalanceado (muchas observaciones negativas por cada una positiva), entrenamos 25 modelos, en el que cada uno incluya dos partes de observaciones negativas por cada una positiva, y se combinan. Se hace 2 - 1 en lugar de 1 - 1 porque se ha observado una mejora en los resultados de esta forma. Tiene sentido, que en caso de duda se “vote” al no.
- Se ha determinado el punto de corte de la probabilidad a partir de la cual se considera el sí en la predicción de la siguiente forma:
 - Se predice utilizando los valores del 0.50 al 0.95 con incrementos de 0.05
 - Se calcula el coste del error, penalizando el falso negativo 4 veces más que el falso positivo.
 - Se escoge la probabilidad con menor coste

Los resultados son los siguientes:

Predicción del nivel de preaviso:

```
##      formula_station prob precision      recall
## 1      N02_28079004 0.95 1.0000000 0.1250000
## 2      N02_28079008 0.55 0.3636364 0.6315789
## 3      N02_28079011 0.55 0.3600000 0.6750000
## 4      N02_28079016 0.75 0.2857143 0.7142857
## 5      N02_28079017 0.55 0.3880597 0.9285714
## 6      N02_28079018 0.90 0.6666667 0.4000000
## 7      N02_28079024 0.75          NA          NA
## 8      N02_28079027 0.95          NA 0.0000000
## 9      N02_28079035 0.90          NA          NA
## 10     N02_28079036 0.85          NA 0.0000000
## 11     N02_28079038 0.65 0.2884615 0.6250000
## 12     N02_28079039 0.50 0.3557692 0.6981132
## 13     N02_28079040 0.75 0.2666667 0.5000000
## 14     N02_28079047 0.85 1.0000000 0.3750000
## 15     N02_28079048 0.90          NA 0.0000000
```

```

## 16    N02_28079050 0.90      NA 0.0000000
## 17    N02_28079054 0.70 0.5084746 0.7142857
## 18    N02_28079055 0.90 1.0000000 0.4000000
## 19    N02_28079056 0.55 0.3229167 0.6739130
## 20    N02_28079057 0.70 0.4035088 0.8518519

```

Predicción del nivel de aviso:

```

##   formula_station prob precision    recall
## 1    N02_28079004 0.90 0.1666667 0.2500000
## 2    N02_28079008 0.60 0.2641509 0.6363636
## 3    N02_28079011 0.70 0.3846154 0.5000000
## 4    N02_28079016 0.90      NA 0.0000000
## 5    N02_28079017 0.65 0.4166667 0.7894737
## 6    N02_28079018 0.90 0.3333333 1.0000000
## 7    N02_28079024 0.85      NA      NA
## 8    N02_28079027 0.90      NA 0.0000000
## 9    N02_28079036 0.80      NA 0.0000000
## 10   N02_28079038 0.95      NA 0.0000000
## 11   N02_28079039 0.55 0.3714286 0.7027027
## 12   N02_28079040 0.85      NA      NA
## 13   N02_28079047 0.85      NA 0.0000000
## 14   N02_28079048 0.80      NA 0.0000000
## 15   N02_28079050 0.90      NA 0.0000000
## 16   N02_28079054 0.70 0.4464286 0.8333333
## 17   N02_28079055 0.85 0.5000000 1.0000000
## 18   N02_28079056 0.80 0.4444444 0.1600000
## 19   N02_28079057 0.70 0.3541667 0.9444444

```

Los valores NA son aquellos en los que el denominador es 0 (true positive + false positive en el caso de la precisión, y true positive + false negative en el recall). Las estaciones que no se listan son aquellas que no tenían ningún caso de preaviso / aviso en el conjunto de validación.

Próximos pasos

El proyecto no termina aquí, los próximos pasos a acometer son:

- Combinación de modelos para mejorar la predicción
- Puesta en producción:
 - Predicción de la intensidad de tráfico a 1 o 2 días
 - Subida del código a un EC2 de Amazon Web Services o similar
 - Lectura del fichero de datos de contaminación en tiempo real cada 30 minutos (cron)
 - Automatización de la predicción y volcado a base de datos
 - Publicación de resultados en un portal: <http://www.contamadrid.es>
- Publicación de resultados en un portal: <http://www.contamadrid.es>
- Extender el análisis a otros contaminantes como el O₃, CO y las PM.
- Pasar el proyecto a un repositorio público (GitHub) por si quieren colaborar terceros